Untying the knot between gestures and speech

Hoetjes, M.W.; Krahmer, E.J.; Swerts, M.G.J.

Published in:
Proceedings of the eight International Conference on Auditory-Visual Speech Processing (AVSP 2009)

Publication date:
2009

Untying the knot between gestures and speech

Marieke Hoetjes, Emiel Krahmer, Marc Swerts
Communication and Cognition group, Humanities Department, Tilburg University, The Netherlands

{M.W.Hoetjes, E.J.Krahmer, M.G.J.Swerts}@uvt.nl

Abstract

Do people speak differently when they cannot use their hands? This study looks at the influence of gestures on speech by having participants take part in an instructional task, half of which had to be performed while sitting on their hands. Other factors that influence the ease of communication, such as visibility and cognitive load, were also taken into account. Results show that lack of visibility or the inability to gesture as well as cognitive load lead to changes in speech and that these factors may influence the successfulness of the instructions.

Index Terms: gesture, visibility, cognitive load, instructional task

1. Introduction

Human interaction is multimodal and, apart from the auditory aspects, the visual aspects of communication such as the gestures people make when they speak can play a large role in the success of this interaction. In fact, a long standing observation in gesture research is that speech and gesture are closely related [1, 2]. However, the exact nature of this relationship is still unclear [as discussed in [3]].

A common viewpoint in most gesture research is that people do not gesture randomly, but that gestures facilitate communication, for the speaker [4, 5], the addressee [6], or both [7]. Given that gestures are an integral component of spoken interactions, one would predict that, if speakers cannot naturally move their hands and/or arms while speaking, this will affect their speech. Moreover, as we will elaborate below, there are reasons to assume that this effect may become more visible when the communicative context becomes more complex, e.g. in situations where speakers cannot see each other as they normally would or when the topic of discussion is a particularly difficult one. Indeed, research has suggested that each of these factors has an influence on gesture production.

When speakers cannot use their entire body in interaction, for example because a speaker is holding an object while speaking, or, in an experimental setting, has to sit on her hands, it becomes difficult, if not impossible, to gesture. Previous research [7-10] has suggested that the close relationship between speech and gesture means that your speech changes when you have to sit on your hands compared to when you are able to gesture. Dobrogaev [8] asked people not to gesture or move their head while speaking and found that speakers' vocabulary size and fluency decreased. However, the details of this study are unclear and unfortunately cannot be recovered. Hostetter, Alibali and Kita [9] immobilized speakers' hands or feet and found that when speakers were unable to use their hands their choice of verbs changed to less rich verbs than when they were able to use their hands. These results are in line with the lexical access hypothesis [11], which states that gestures facilitate speech, and preventing people from gesturing will cause speech to become less fluent. However, apart from the influence on verb choice, no other effects of gesture prevention on speech production were found in Hostetter, Alibali and Kita's study. Emmorey and Casey [7] looked at speech during gesture prevention and lack of visibility between speakers and found that gesture and speech complement each other, but this study only focused on the use of spatial language. Krahmer and Swerts [10] showed that the movement of producing beat gestures has an influence on the prosodic properties of the co-occurring speech. However, Krahmer and Swerts' study [10] used an artificial task and their results therefore do not necessarily generalize to natural speech.

A communicative context can be more complex than usual when there is no mutual visibility. Previous studies [7, 12-14] have looked at the influence of mutual visibility on the production of gestures and have found that speakers still gesture when they cannot see their addressee, although the exact nature of their gestures changes. When people cannot see each other, their gestures tend to be smaller than when they can see each other [12, 14] and the number of gestures performed decreases [12, 13]. The fact that gestures are still produced when there is no visibility between speakers, however, suggests that the link between speech and gesture is not only based on communicative aspects such as visibility but that gestures serve some speaker internal purpose as well.

Another example of a way in which communication can be more complex than normal is when speakers have to perform tasks of differing complexity. Several studies suggest different results with regard to gestures and their relationship to complex speech. On the one hand, as mentioned above, it has been argued that gestures facilitate lexical access [4, 11] and are thus mainly produced for the speaker herself. More complex tasks and a larger cognitive load will thus lead to more gestures. On the other hand, research has also suggested that gestures are largely produced for the addressee and thus serve a communicative purpose [6, 13], meaning that the number of gestures should stay the same regardless of the difficulty of the speaker's task. Recent research [15] shows that these findings do not have to contradict each other but that task difficulty does have an effect on the frequency of gesture use.

A speaker might have to deal with an increased cognitive load due to task complexity, but the cognitive load may also be decreased because the subject of discussion has already

been dealt with before. Research on second language acquisition [14, 16] has shown that speakers continue to produce gestures when a subject has already been dealt with before. Research on native speakers [17] has shown that a decreased cognitive load due to common ground can lead to a higher gesture rate.

To sum up, previous research has shown the strength of the relationship between speech and gesture, by looking at the influence of the (in)ability to gesture on speech, the influence of mutual visibility on gesture production and the influence of task complexity and cognitive load on gesture production. However, many of the previous studies have only looked at certain aspects of speech or gesture and it is unclear how these various potentially relevant factors of complexity can be related to each other. Moreover, because of the artificial nature of some of the tasks used, we do not know whether the findings can be generalized to other settings. This means that many aspects of the direct influence of gestures on speech remain unknown.

In the present study, an integrated naturalistic approach is taken to study the influence of gestures on speech in a new experimental paradigm. The previous findings are tested by looking at natural speech in several communicatively difficult situations, namely when subjects had to sit on their hands, when there was no mutual visibility and during tasks with differing cognitive load. The experiment takes place in the form of a tie-knotting instructional task, which will combine speech that is as natural as possible with a setting in which it can be expected that speakers will gesture. The task allows for measurement of successful instructions by looking at the end result of the tie knot and enables the manipulation of ability to gesture, mutual visibility and cognitive load. Data analysis is ongoing, but preliminary results will be presented below.

2. Method

The goal of this study is to compare speech when people are unable or unlikely to gesture to speech when people are able and likely to gesture, i.e. when they can use their entire body and see their addressee. An experimental paradigm was developed in which participants took part in an instructional task where one of the participants, the instructor, had to watch video clips depicting a person tying different kinds of tie knots and instruct the other participant, the matcher, to tie a tie in the same manner as in the video clips.

2.1. Participants

In total, 38 pairs of native speakers of Dutch took part (25 male participants, 51 female participants). Each pair consisted of one instructor and one matcher. The participants were mainly first year university students (mean age 20 years old, range 17-32 years old) who, in some cases, knew each other. Participants took part in random pairs (these could be male, female, or mixed pairs). Participants took part in the experiment as partial fulfilment of course credits.

2.2. Stimuli

Instructors watched two different video clips on a laptop, containing instructions on how to tie two different types of tie knot. To control for cognitive load, two tie knots were chosen which have been shown to differ in complexity (the easy Kelvin tie knot and the more complicated Persian tie knot). Each clip with one type of tie knot instruction was presented three times before the other clip was presented three times. Each separate clip, containing instructions for a different tie knot, was cut into six fragments. Each fragment contained a short (maximally 10 or 15 seconds) instructional step for the knotting of a tie. The clips contained the upper body of a person who slowly knotted a tie without speaking or using facial expressions. Each fragment was accompanied by a small number of key phrases, such as '...wide...under...thin...', 'tight' or '...through...loop...' (these key phrases were given in Dutch). A still from one of the clips' fragments can be seen in figure 1.

Figure 1. Still of the beginning of a fragment of one of the stimulus clips, in this case accompanied by the phrases 'behind' and 'up'.

2.3. Setup

The participants entered the lab in pairs and were randomly allocated the role of instructor or matcher. The two participants sat down in seats that were positioned opposite each other. The seat of the instructor did not contain any armrests. Participants were asked to sign a consent form, were given instructions and the possibility to ask for clarifications, after which the experiment would start.

Instructors then watched all six fragments of one tie knot on a laptop and gave instructions to the matcher after watching each fragment. The matcher could not see the screen of the laptop. This procedure was repeated three times for the same tie knot, after which the fragments for the other tie knot were shown three times. Matchers thus had to tie the same tie knot three times followed by the other type of tie knot which also had to be tied three times. After the experiment, participants filled out a short questionnaire, asking, among other things, about their experience with tie knotting and whether they knew the person they had just done the experiment with. The experiment took about 30 minutes.

Half of the instructors had to sit on their hands for the first half of the experiment, whereas the other half of the instructors had to sit on their hands during the second half of the experiment. This means that all instructors conducted half of the task while sitting on their hands.

For half of all participant pairs, an opaque screen was placed in between the instructor and the matcher.

The order in which instructors were presented with the two different tie knots was counterbalanced over participants.

The experimenter was in the lab during the experiment and controlled the speed with which the video fragments were shown. This was due to the fact that the instructors were unable to control the laptop while they were sitting on their hands. The experimenter switched to the next fragment when it was clear that the instructor had said all there was to say and the matcher had understood the instructions and tied (part of) the knot accordingly.

The proceedings of the experiment were videotaped (audio and video) and after each entire video clip (consisting of six fragments) had been watched and instructed, a picture was taken by the experimenter of the end state of the tied tie knot on the matcher.

2.4. Design

The experiment was a mixed design model, with one between subjects factor, namely whether there was a screen between participants or not and three within subjects factors, namely whether instructors were able to gesture or not, type of tie knot (2 levels) and number of times the clip had been presented (3 levels).

Half of the pairs had a screen between them for the entire duration of the experiment, the other half were able to see each other during the experiment. Instructors had to sit on their hands either during the first half of the experiment or during the second half of the experiment (this order was counterbalanced). All instructors had to instruct the two different tie knots three times.

2.5. Data analysis

Several types of data were collected from this experiment. Video data from the instructor was recorded (including low-quality audio from the video camera), high quality audio data from both the instructor and the matcher was recorded, and a total of six photographs of the end state of the tie knot on each matcher were taken.

The video data has been analysed with regard to the length of the instructor's speech, both in time and in number of words, across conditions. Initial gesture analysis has identified all gestures produced by the instructor. The photographs of the end state of each tie knot on the matcher have been judged according to the correctness of the tie knot. These analyses were chosen because it could be expected that instructors would take longer, either in time or in number of words, to instruct the tie knot when the communicative situation is more difficult than it normally is, for example because of the inability to gesture, because of limited visibility, or because of a large cognitive load. The number of gestures produced by the instructor can tell us whether, as in previous research [7, 12-14], gestures are still produced when people cannot see each other. The photographs of the end state of the tie knots were coded and these were taken into account because the end state of the tie knot can be seen as a measurement of the successfulness of the instructions.

3. Results

Data analysis is preliminary and ongoing, but a first few explorative results can be discussed. The video data was annotated in a multimodal annotation programme, ELAN [18], looking at the length in time between two fragments, the number of words used and the gestures produced in (part of) the instructor's speech. The photographs of the end state of each tie knot were coded by six independent judges according to the correctness of the tie knot.

3.1. Length of speech in time

The length of speech in time in seconds was measured between the start of one video clip fragment and the start of the following video clip fragment. Because of the inability of the instructors to use their hands during part of the experiment, the experimenter controlled the beginning of each video clip fragment. This was done by making sure the instructor had given all instructions and the matcher had tied part of the tie knot before going on to the next fragment. The mean length of all fragments was 31 seconds (*SD*=12 seconds). There were no significant differences between the length of speech when the instructor could gesture and the length of speech when the instructor had to sit on her hands. There were also no significant differences between the length of speech of instructors who could see their addressee and the length of speech of instructors who could not see their addressee. There were, however, significant differences in the length of speech in time between fragments with a difference in cognitive load, namely the type of tie knot, $F(1,60)= 5.523$, $p<.05$, and the number of times the tie knot has been presented, $F(2,59)= 15.774$, $p<.05$ (see table 1 and 2). The easier tie knot takes less time to instruct and people get quicker in instructing a tie knot when they have done it before.

Table 1. Mean length of one fragment of each tie knot in seconds.

Kelvin tie knot (*SD*)	Persian tie knot (*SD*)
27.9 (1.9)	34.0 (1.9)

Table 2. Mean length of one fragment in seconds, for each number of times the tie knot has been instructed.

First (*SD*)	Second (*SD*)	Third (*SD*)
35.7 (1.5)	30.2 (1.4)	27.1 (1.6)

3.2. Length of speech in words

Presently, one third of the speech of the instructor has been transcribed verbatim. Instructions given for the first attempt for each tie knot were transcribed. Since all the instructors had to sit on their hands for half of the experiment, this means that for each instructor, one tie knot instruction (consisting of instructions for all six fragments) was transcribed when she was allowed to use her hands and one tie knot instruction was transcribed when she had to sit on her hands. This led to a total of 456 transcribed fragment instructions, half of which was produced when there was a screen between participants and half of which was produced when participants could see each other. The number of words for each of these instructions was counted, including filled pauses (e.g. 'uhm') and comments about the experiment itself (e.g. 'can I see the clip again?').

No main effects of visibility or ability to gesture were found. However, as can be seen in table 3, there was a two-way interaction effect between visibility and ability to gesture. Significantly fewer words were used to instruct a fragment, $F(1,452)= 4.249$, $p<.05$, when the instructor was able to use her hands *and* was able to see the matcher than when the instructor was not able to see the matcher or was unable to use her hands. When instructors were not able to gesture, the number of words used was the same, regardless of whether participants were able to see each other or not. However, when instructors were able to gesture, mutual visibility had an effect on the number of words used, with instructors who did not see the matcher (and where consequently the matcher did not see the instructor and her gestures) using more words than instructors who did see the matcher.

Table 3. Mean number of words used in one instructional fragment, in one third of the data.

Able to gesture / Mutual visibility	Yes	No	Mean total
Yes	36	50	43
No	51	49	50
Mean total	44.5	49.5	46.5

3.3. Number of gestures

The part of the data that was transcribed verbatim (one third of all the data) has also been analysed with regard to the number of gestures produced by the instructor. As can be seen in table 4, an unsurprising, significant, difference was found between the mean number of gestures each instructor produced when instructors had to sit on their hands ($M= .71$) and the mean number of gestures produced when instructors were free to use their hands ($M= 12.68$, $F(1,72)= 27.056$, $p<.001$). No other significant differences were found. Noteworthy however, is the fact that instructors do still gesture even though they have to sit on their hands and that instructors still gesture frequently when there is a screen between themselves and the matcher.

Table 4. Mean number of gestures produced by each instructor, in one third of the data .

Able to gesture / Mutual visibility	Yes	No	Mean total
Yes	14.42	.37	7.4
No	10.95	1.05	6
Mean total	12.68	.71	6.7

3.4. Quality of tie knots

Six photographs were taken of each matcher as soon as he or she had finished tying the tie knot. These photographs (examples of which can be seen in figure 2 and figure 3) were presented to six independent judges who saw all the printed photographs and had to score the tie knots on a range from 1 to 5, with a tie knot of a more or less perfect quality getting a score of 5 and a completely incorrect tie knot getting a score of 1.

Photographs of tie knots taken when there was a screen between participants ($M=2.35$) were scored significantly lower than photographs of tie knots taken when there was no screen between participants ($M=2.65$), $F(1,226)=4.8$, $p<.05$. It seems that visual feedback between participants leads to a better tie knot. There were no significant differences between judgments of photographs of tie knots taken when the instructor had to sit on her hands and judgments of photographs of tie knots taken when the instructor was free to move her hands.

Figure 2. Example of photograph of tie knot made when there was no screen between speakers.

Figure 3. Example of photograph of tie knot made when there was a screen between speakers.

4. Discussion

Although the analyses are ongoing, the different types of data have already shown some interesting and different results.

The length of speech in time has shown that instructors take longer when they are instructing a difficult tie knot compared to the easier tie knot and that they get quicker once they have instructed the same tie knot before. The type of tie knot and the version of the instruction can be qualified as aspects of cognitive load. Cognitive load seems to have an effect on the time people need to give an instruction.

The results for the length of speech in number of words show an interaction effect of visibility and ability to gesture, with fewer words needed by the instructor to instruct a

fragment in a completely natural situation (with mutual visibility and ability to gesture) than in a situation with lack of mutual visibility or inability to gesture. As in previous research [7-9, 12-14], visibility and ability to gesture have an effect on speech. It is not yet clear whether the number of words used also depends on cognitive load, with the possibilities that more complex tie knots might require more words or the same number, but a different type of words.

The results from the mean number of gestures produced by the instructors have shown that instructors, unsurprisingly, gesture significantly less when they are not able to gesture than when they are able to gesture, but that the inability to gesture does not mean that people do not gesture at all. Also, lack of mutual visibility does not lead to significantly fewer gestures. It remains to be seen whether the number of gestures produced changes depending on cognitive load (looking at the different tie knots and the number of times they have been instructed before) and whether this is related to visibility as in previous research [15].

The analyses of the judgment scores given to the photographs of the tie knots have shown that tie knots tied when participants could see each other were judged to be of better quality than tie knots tied by participants who could not see each other. Apart from the explanation that this may be due to the lack of visual feedback when participants could not see each other it might also be the case that the less well tied ties were a result of different instructions in the case where participants could not see each other compared to when participants could see each other. Further analysis of the speech data would have to confirm this.

Apart from expanding the present analyses of which some results have been given above, many more aspects of speech can be taken into account when looking at the data collected in this experiment. Semantic analyses of the words used by the instructor can be used to see whether for example more filled pauses (e.g. 'uhm') are used in communicatively difficult situations or whether less rich verbs are used when the instructor cannot gesture, as in Hostetter, Alibali and Kita [5]. Prosodic analyses of the instructors' speech can shed light on possible differences in pitch height and intonational contours of the instructor between conditions. It might for example be the case that instructors make more use of intonation when they cannot use their hands or when the matchers cannot see them. It could, however, also be the case that the instructors' speech becomes more monotonous when they cannot use their hands, as in Dobrogaev [8], since gestures have been found to have an influence on prosodic prominence [10].

This study has taken the view that it is possible to untangle the relationship between speech and gesture by looking at what happens in speech in the absence of gestures and the presence of gestures in this dataset has presently been given little attention. Questions that may be answered by looking more closely at the number and type of gestures performed are whether people are more or less likely to gesture when they have to perform a difficult task and whether the timing of these gestures differs depending on the condition in which they are produced. Also, initial results have already shown the strength of the relationship between speech and gesture since in quite a few cases, instructors produced gestures even when they had to sit on their hands, often accompanied by an apology to the experimenter. The timing of these 'slips of the hand' could provide us with more information about situations in which it is deemed absolutely necessary to use a gesture.

5. Conclusions

Despite the fact that only a third of the data has been analysed presently, we can already say that the inability to gesture by the instructor and the inability to see gestures by the matcher have an influence on the instructors' speech and on the successfulness of the interaction between instructor and matcher. It is not yet clear whether the inability to gesture has a larger influence on speech than the lack of visibility between speakers or the level of cognitive load. Further analyses will have to show whether it is a tie between these factors or whether gestures have such a large influence on speech that when speakers are tied down they get tied up in knots.

6. Acknowledgements

We would like to thank Bas Roset for help in creating the stimuli and Joost Driessen for help in annotating the data. We received financial support from The Netherlands Organization for Scientific Research, via a Vici grant (NWO grant 277-70-007), which is gratefully acknowledged.

7. References

[1] A. Kendon, "Some reasons for studying gesture," *Semiotica,* vol. 62, pp. 3-28, 1986.

[2] D. McNeill, *Hand and mind. What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[3] J. P. de Ruiter, "The production of gesture and speech," in *Language and gesture*, D. McNeill, Ed. Cambridge: Cambridge University Press, 2000, pp. 284-311.

[4] R. M. Krauss and U. Hadar, "The role of speech-related arm/hand gestures in word retrieval," in *Gesture, speech, and sign*, R. Campbell and L. Messing, Eds. Oxford: Oxford University Press, 2001, pp. 93-116.

[5] M. Alibali, S. Kita, and A. Young, "Gesture and the process of speech production: We think, therefore we gesture," *Language and cognitive processes,* vol. 15, pp. 593-613, 2000.

[6] A. Özyürek, "Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures," *Journal of Memory and Language,* vol. 46, pp. 688–704, 2002.

[7] K. Emmorey and S. Casey, "Gesture, thought, and spatial language," *Gesture,* vol. 1, pp. 35-50, 2001.

[8] S. M. Dobrogaev, "Ucnenie o reflekse v problemakh iazykovedeniia [Observations on reflexes and issues in language study]," *Iazykovedenie i Materializm* pp. 105-173, 1929.

[9] A. B. Hostetter, M. W. Alibali, and S. Kita, "Does sitting on your hands make you bite your tongue? The effects of gesture prohibition on speech during motor descriptions," in *Proceedings of the 29th annual meeting of the Cognitive Science Society*, D. S. McNamara and J. G. Trafton, Eds. Mawhah, NJ: Erlbaum, 2007, pp. 1097-1102.

[10] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: acoustic analyses, auditory preception and visual perception," *Journal of Memory and Language,* vol. 57, pp. 396-414, 2007.
[11] F. H. Rauscher, R. M. Krauss, and Y. Chen, "Gesture, speech and lexical access: The role of lexical movements in speech production," *Psychological Science,* vol. 7, pp. 226-230, 1996.
[12] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost, "Gesturing on the telephone: Independent effects of dialogue and visibility," *Journal of Memory and Language,* vol. 58, pp. 495–520, 2008.
[13] M. Alibali, D. C. Heath, and H. J. Myers, "Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen," *Journal of Memory and Language,* vol. 44, pp. 169–188, 2001.
[14] M. Gullberg, "Handling discourse: gestures, reference tracking, and communication strategies in early L2," *Language Learning,* vol. 56, pp. 155-196, 2006.
[15] L. Mol, E. Krahmer, A. Maes, and M. Swerts, "Gesturing and cognitive load," in *CogSci 2009*, Amsterdam, 2009.
[16] K. Yoshioka, "Gesture and information structure in first and second language," *Gesture,* vol. 8, pp. 236-255, 2008.
[17] J. Holler and K. Wilkin, "Communicating common ground: how mutually shared knowledge influences speech and gesture in a narrative task," *Language and cognitive processes,* vol. 24, pp. 267-289, 2009.
[18] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research.," in *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 2006.